Datenaufbereitung und Dokumentation

Brislinger, Evelyn; Moschner, Meinhard

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Auszug aus dem Buch:

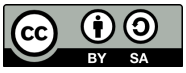
Uwe Jensen
Sebastian Netscher
Katrin Weller (Hrsg.)

Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten

Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten

Verlag Barbara Budrich
Opladen • Berlin • Toronto 2019

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.



Dieses Buch steht im Open-Access-Bereich der Verlagsseite zum kostenlosen Download bereit (https://doi.org/10.3224/84742233).
Eine kostenpflichtige Druckversion (Print on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-8474-2233-4 (Paperback)
eISBN 978-3-8474-1260-1 (eBook)
DOI 10.3224/84742233

Umschlaggestaltung: Bettina Lehfeldt, Kleinmachnow – www.lehfeldtgraphic.de
Lektorat: Nadine Jenke, Potsdam
Satz: Anja Borkam, Jena – kontakt@lektorat-borkam.de
Titelbildnachweis: Foto: Florian Losch
Druck: paper & tinta, Warschau
Printed in Europe

6. Datenaufbereitung und Dokumentation

Evelyn Brislinger und Meinhard Moschner

6.1 Datenaufbereitung im Lebenszyklus eines Projekts

Die Datenaufbereitung und Dokumentation als eine Phase im Verlauf eines empirischen Forschungsprojekts hat das Ziel, die erhobenen Daten für die Forschung nutzbar zu machen. Hierbei werden die Daten codiert, überprüft, bearbeitet und dokumentiert. Das stellt Forschende vor die Aufgabe, die einzelnen Arbeitsschritte zu definieren und zu einem Workflow zusammenzusetzen. Dieser geht idealerweise von der Art und Komplexität der Daten aus und hilft, die Ziele des Projekts umzusetzen, ohne die oft begrenzten zeitlichen und finanziellen Ressourcen aus dem Blick zu verlieren.

Unter empirischen Forschungsprojekten verstehen wir in diesem Kapitel Projekte, die auf der Grundlage eines Erhebungsinstruments einfache oder komplexe zeit- und/oder ländervergleichende Daten erheben und diese als Analysefiles aufbereiten. Die Projektziele können ausgehend von den Prinzipien guter wissenschaftlicher Praxis zunächst auf die Datenqualität, im Sinne möglichst fehlerfreier Daten, gerichtet sein. Darüber hinausgehend können sie weitere Möglichkeiten der Datennutzung, wie z.B. die Replikation der Projektergebnisse oder auch eine Nachnutzung der Daten durch Dritte, eröffnen. Mit der Komplexität der Daten und dem Wunsch, den Prozess ihrer Entstehung und Bearbeitung transparent zu machen, wachsen gleichzeitig die Anforderungen an ihre Aufbereitung und Dokumentation. Vorgenommene Datenmodifikationen müssen dann auch für Forschende außerhalb des Projekts nachvollziehbar und Datenprobleme gut dokumentiert sein. Nur so bleibt die Datenqualität für nachfolgende Analysen bewertbar und es erschließen sich insbesondere komplexe Datenfiles auch ohne internes Projektwissen.

Die Angebote von Repositorien und Datenarchiven können dann genutzt werden, um die Projektergebnisse nachhaltig zu sichern und für eine erneute Nutzung bereitzustellen (DFG 2015). Hierfür muss am Ende eines Projekts eine Dokumentation zur Verfügung stehen, die Datennutzenden eine Bewertung des Inhalts und der Qualität der Daten ermöglicht und gleichzeitig den Standards des gewählten Repositoriums oder Datenarchivs entspricht. Erfahrungsgemäß gelingt dies eher dann und ohne zusätzliche Ressourcen, wenn dieser Schritt frühzeitig geplant wird und die Informationen und Dokumentationen im Projektverlauf zeitnah aufgebaut und systematisch organisiert werden (Ball 2012: 3).

Da im Alltag empirischer Forschungsprojekte die Analyse der erhobenen Daten und die Publikation der Forschungsergebnisse im Vordergrund stehen, sind Ressourcen und Zeit für ihre tiefere Aufbereitung und umfassende Dokumentation oft begrenzt. Projekte stehen demzufolge vor der Herausforderung, einen Datenaufbereitungsworkflow zu entwickeln, der möglichst schlank ist und gleichzeitig eine hohe Datenqualität sowie umfassende Dokumentation ermöglicht. Ausgehend hiervon richtet sich der Beitrag auf die Frage, wie ein systematisches Forschungsdatenmanagement (Jensen 2011) helfen kann, die Ziele eines Projekts in praxisnahe Workflows und Arbeitsschritte zu übersetzen und diese zwischen den beteiligten Personen zu kommunizieren.

Hierfür erörtern wir zunächst die Bedeutung der Planung der Datenaufbereitungsschritte im Lebenszyklus eines Forschungsprojekts (Kapitel 6.2). Wir gehen dann auf Regeln, Standards und Prozeduren ein, die für die Variablencodierung (Kapitel 6.3) sowie für die Prüfung

https://doi.org/10.3224/84742233.07

und Behandlung von Datenfehlern (Kapitel 6.4), als Kernaufgaben der Datenaufbereitung, erforderlich sind. Auf dieser Grundlage entwickeln wir das Modell eines Datenaufbereitungsworkflows (Kapitel 6.5), der die geplanten Bearbeitungsschritte zusammenfasst und hilft, diese in die Praxis umzusetzen. Abschließend erörtern wir die Bedeutung des Transfers der Daten und Informationen im Projektverlauf für den Aufbau einer umfassenden Dokumentation (Kapitel 6.6).

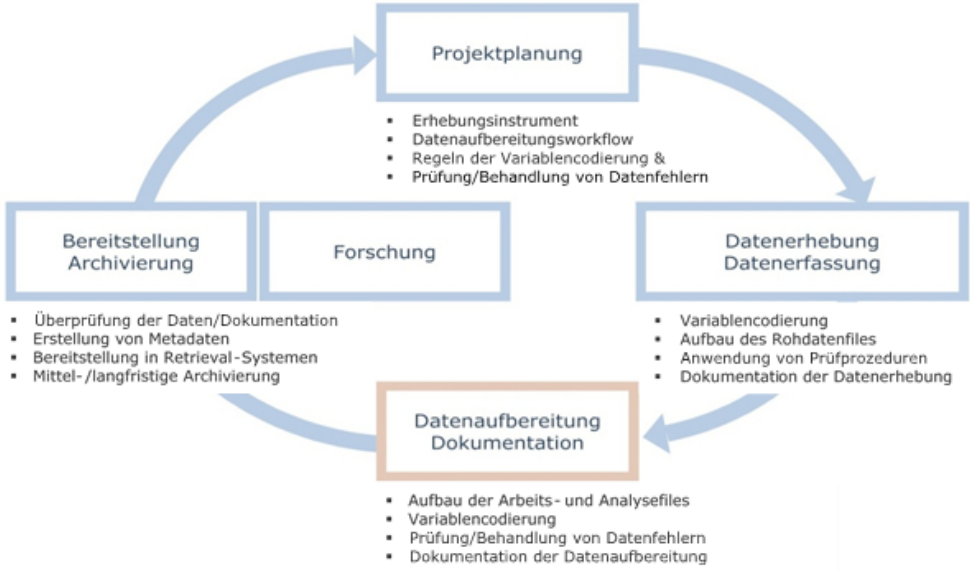
Die Beispiele, die wir hierfür verwenden, stammen z.T. aus größeren Forschungsprojekten. Sie illustrieren den Grundgedanken des Beitrags, zu Beginn eines Projekts Datenaufbereitungsregeln zu vereinbaren und die erforderlichen Datensätze und Variablen zu definieren, hieraus geeignete Bearbeitungsschritte abzuleiten und schließlich die einzelnen Elemente zu einem Workflow zusammenzusetzen. Sie lassen sich somit gleichermaßen in die Praxis mittlerer und kleinerer Projekte umsetzen.

6.2 Planung der Datenaufbereitung und Dokumentation

Lebenszyklusmodelle stellen ein hilfreiches Werkzeug für die Planung und das Management von Forschungsdaten dar, indem sie es ermöglichen, die erforderlichen Schritte der Entstehung und Bearbeitung der Daten frühzeitig zu bedenken (Ball 2012: 3). Eine umfassende Darstellung des Lebenszyklus von Forschungsdaten findet sich in Kapitel 2.2. Abbildung 6.1 illustriert ein vereinfachtes Modell, auf das wir in diesem Kapitel in verschiedenen Zusammenhängen zurückkommen werden. Beim Aufbau eines Lebenszyklusmodells gehen Forschende idealerweise von der Art, den Zielen und der Komplexität des Projekts aus und definieren die erforderlichen Projektphasen. Für die einzelnen Phasen können dann, wie in der Abbildung exemplarisch gezeigt, die Arbeitsschritte sowie die hierbei zu generierenden Daten und Informationen geplant werden. Auf dieser Grundlage wiederum werden die Verantwortlichkeiten zwischen den am Projekt beteiligten Akteuren vereinbart und der Transfer der Daten und Informationen zwischen ihnen organisiert. Schließlich kann das entstandene Lebenszyklusmodell hinsichtlich redundanter Arbeitsschritte, wie z.B. wiederholt eingesetzte Prozeduren der Datenbearbeitung, überprüft und ggf. vereinfacht werden. Für eine umfassende und detaillierte Beschreibung der Phasen empirischer Forschungsprojekte verweisen wir auf die Cross-Cultural Survey Guidelines (Survey Research Center 2016).

Schaut man auf die Daten, die die Phasen des Lebenszyklusmodells durchlaufen, zeigt sich ein einfacher Workflow: In der Planungsphase wird das Erhebungsinstrument entwickelt. Auf dieser Grundlage werden die Daten erhoben und in das sogenannte Rohdatenfile überführt. Die Rohdaten werden weiter bearbeitet und als interne Arbeitsfiles gesichert. Diese wiederum bilden den Ausgangspunkt für die Generierung der Analysefiles, die für eine Nutzung innerhalb und außerhalb des Projekts bereitgestellt werden können. Bei kleinen Projekten sind an diesem Prozess das Projektteam und ggf. ein Datenarchiv oder Repositorium beteiligt. Mittlere und große Projekte arbeiten darüber hinaus oft mit einem oder mehreren Erhebungsinstituten bzw. mit regional verteilten Projektpartnern zusammen.

Abbildung 6.1: Datenaufbereitung als Phase im Lebenszyklus von Forschungsdaten



Quelle: Eigene Darstellung in Anlehnung an Survey Research Center (2016)

Betrachtet man nur die Phase der *Datenaufbereitung und Dokumentation* im Lebenszyklusmodell, dann müssen hierfür Regeln, Standards und Prozeduren vereinbart werden. Diese umfassen den Prozess von der Definition der einzelnen Variablen bis hin zur Bereitstellung des Analysefiles für die Forschung. Bei der Planung der erforderlichen Arbeitsschritte und Ressourcen sollten Faktoren, wie die Komplexität und Fehleranfälligkeit der Daten, die angezielte Datenqualität sowie die erwartete Datennutzung berücksichtigt werden (Lück/Landrock 2014: 405; Jensen 2012: 16f.).

So erfordern komplexe Daten, die z.B. im Ergebnis der Kumulation von Erhebungswellen oder der Integration nationaler Daten entstehen, eine umfassendere Dokumentation. Gleichermaßen lassen komplizierte Filterführungen oder eine Reihe offener Fragen im Erhebungsinstrument einen höheren Aufwand für die Variablencodierung und Überprüfung der Daten erwarten. Demgegenüber kann der Aufwand für die Datenaufbereitung verringert werden, wenn die Antworten der Befragten bereits während der Datenerhebung standardisiert und überprüft werden können. Darüber hinaus ermöglichen die Dokumentation der einzelnen Prüfschritte sowie der Transfer der Prüfergebnisse im Projektverlauf, nachfolgende Bearbeitungsschritte gezielt vorzunehmen und redundante Datenmodifikationen einzuschränken. Das wiederum kann die Gefahr neuer Datenfehler, die generell bei der Überprüfung und Bearbeitung der Daten entstehen können, oder auch die Gefahr einer verzögerten Bereitstellung der Analysefiles für die Forschung verringern (Survey Research Center 2016: 654). Um eine gute Balance zwischen der Datenqualität, der Projektzeit und den Ressourcen finden zu können, sollte demzufolge bereits zu Projektbeginn entschieden werden, welche Arbeitsschritte in welchen Phasen des Projekts durchzuführen sind (Schäfer et al. 2006: 1).

Im Folgenden geben wir eine kurze Beschreibung der in Abbildung 6.1 beschriebenen Phasen des Lebenszyklusmodells und setzen hierbei den Schwerpunkt auf ihre besondere Bedeutung für die *Datenaufbereitung und Dokumentation*.

6.2.1 Projektplanung: Projektempfehlungen und Erhebungsinstrument

Geht man von Best-Practice-Beispielen, wie z.B. dem European Social Survey (ESS)[1], aus, dann werden in der *Planungsphase* die Regeln, Standards und Prozeduren für die folgenden Phasen der *Erhebung*, *Aufbereitung und Dokumentation* sowie der *Bereitstellung* der Daten ausgearbeitet. Das Projektteam entwickelt das Erhebungsinstrument und legt damit bereits die Grundlage für die nachfolgenden Schritte der Codierung und Überprüfung der Variablen. Darüber hinaus wird festlegt, welche Daten und Dokumentationen produziert werden, welche Regeln für die Codierung, Überprüfung, Bearbeitung und Dokumentation der Daten anzuwenden sind und welche projektspezifischen bzw. internationalen Standards genutzt werden sollen (Kolsrud et al. 2010). In Kooperation mit einem Datenarchiv werden geeignete Publikationskanäle für die Daten vereinbart, um ggf. erforderliche Anonymisierungsschritte frühzeitig bedenken zu können. Die hierbei entstehenden Projektempfehlungen haben das Ziel, den Entstehungs- und Aufbereitungsprozess der Daten transparent zu machen und vorgenommene Datenmodifikationen auch nach Abschluss des Projekts nachvollziehen zu können (vgl. GESIS Survey Guidelines (GESIS 2016); Guidelines for Best Practice in Cross-cultural Surveys (Survey Research Center 2016)).

6.2.2 Datenerhebung: Rohdatensatz und Dokumentation

In der *Erhebungsphase* werden die Antworten der Befragten erhoben und zumeist durch ein Erhebungsinstitut in ein Datenformat gebracht, das der Struktur des Erhebungsinstruments entspricht. Je nach Erhebungsmethode werden die Antworten in digital programmierten Fragebögen erfasst (*computer-assisted interview*, CAI) oder von Hand in gedruckten Fragebögen eingegeben (*paper and pencil interview*, PAPI bzw. schriftliche Befragung). Letzteres macht einen weiteren Schritt der digitalen Erfassung der Antworten der Befragten erforderlich. In kleineren Projekten kann dieser Schritt durch die Erstellung einer Datenmatrix und die manuelle Dateneingabe mit Hilfe von Dateneditoren erfolgen. Hilfreiche Tools werden von Statistiksoftware wie SPSS, STATA oder SAS bereitgestellt (Lück/Landrock 2014: 400f.; Jensen 2012: 23). Im Ergebnis dieser Phase entsteht das Rohdatenfile. Die Daten werden zusammen mit methodischen Informationen über den Erhebungs- und Erfassungsprozess und ggf. Paradaten[2] an die Phase der Datenaufbereitung übergeben.

6.2.3 Datenaufbereitung: Arbeits- und Analysefiles und Dokumentation

Während der *Datenaufbereitung* werden die Rohdaten weiter codiert, überprüft, bereinigt und dokumentiert. Ein Projekt kann sich hierbei das Ziel stellen, möglichst fehlerfreie und gut dokumentierte Daten zu generieren bzw. darüber hinausgehend die Daten durch weitere Schritte der Standardisierung und Harmonisierung aufzuwerten. So werden z.B. die Einzeldatensätze, die bei zeit- oder ländervergleichenden Erhebungen entstehen, weiter technisch standardisiert und inhaltlich harmonisiert. Die Verwendung einheitlicher Variablennamen und Werte für die Fragen im Erhebungsinstrument bzw. die Harmonisierung der zeit- oder

1 Der European Social Survey (ESS) ist eine auf wissenschaftlichen Standards beruhende länderübergreifende Erhebung, die seit 2001 alle zwei Jahre europaweit durchgeführt wird.

2 Bei Paradaten handelt es sich um Daten, die im Prozess der Erhebung von Umfragedaten entstehen. Je nach Erhebungsmodus sind dies z.B. die Anzahl und das Ergebnis der Kontaktaufnahmen mit den zu befragenden Personen, der Zeitpunkt und die Dauer der Befragung, die Beschreibung der Wohngegend bzw. des Haushalts ebenso wie Anmerkungen zum individuellen Antwortverhalten (vgl. Felderer/Birg/Kreuter 2014).

länderspezifisch erhobenen Variablen ermöglicht ihre anschließende Integration in einen Gesamtdatensatz. Gleichermaßen können Datentransformationen vorgenommen werden, um die Nutzbarkeit der Daten sowie ihre Vergleichbarkeit innerhalb oder zwischen Datenkollektionen zu verbessern. Die in dieser Phase entstehenden internen Master-Arbeitsfiles sind das Ergebnis aller vorgenommenen Bearbeitungsschritte und bilden die Grundlage für die zu generierenden Analysefiles. Diese stellen eine zumeist reduzierte Datensatzversion dar, in der z.B. interne Hilfsvariablen gelöscht und die Daten für eine Nachnutzung durch Dritte stärker anonymisiert sind. Die Datenfiles werden zusammen mit der Dokumentation der vorgenommenen Arbeitsschritte und Datenmodifikationen für eine nachhaltige Sicherung bzw. breite Nutzung an ein Datenarchiv oder Repositorium weitergegeben.

6.2.4 Bereitstellung: Analysefiles, Dokumentation und Metadaten

Datenarchive oder Repositorien, die am Ende des Lebenszyklusmodells stehen, können naturgemäß nur die Daten und Informationen für eine breitere Nutzung zur Verfügung stellen, die im Projektverlauf generiert wurden und nach Projektabschluss noch verfügbar sind. Demgegenüber stehen Projekte oft vor der Herausforderung, große Mengen an Dateien organisieren bzw. die Verluste an Projektwissen infolge des Wechsels im Projektteam kompensieren zu müssen. Eventuelle Versäumnisse im projektinternen Forschungsdatenmanagement zeigen sich spätestens bei der Zusammenstellung der Daten und Dokumentation, die für eine breitere Nachnutzung erforderlich sind. Entsprechen diese nicht den Qualitätserfordernissen und Standards z.B. eines Datenarchivs oder wurden Datenschutzerfordernisse nicht eingehalten, können auch nach Projektabschluss noch Arbeitsschritte erforderlich werden.

6.3 Schritte der Variablencodierung

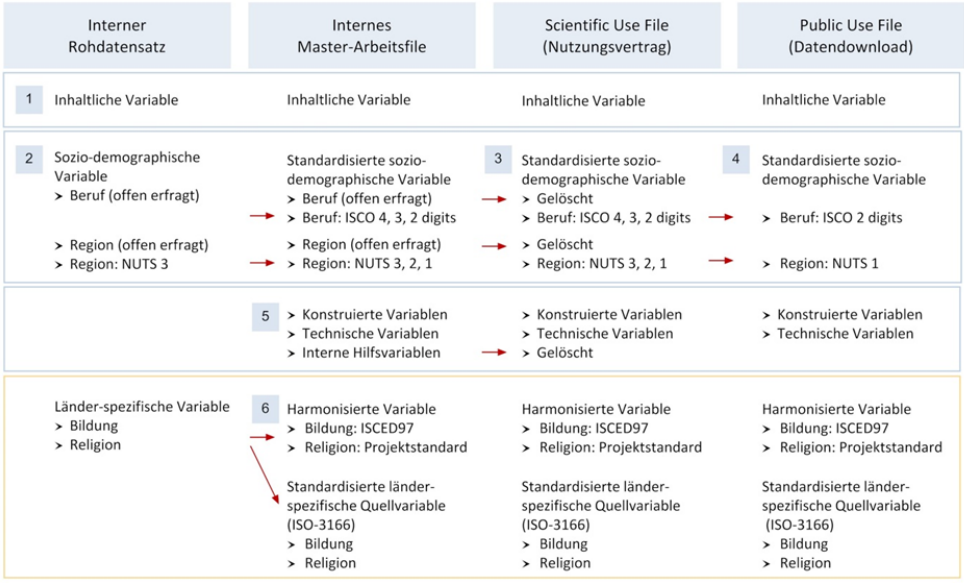
Sind die im Projektverlauf aufzubauenden Datenfiles geplant, kann im Weiteren vereinbart werden, welche Variablen sie enthalten sollen und welche Codierschritte hierfür erforderlich sind. Das ist die Aufgabe der Variablencodierung, in deren Ergebnis Datensätze mit eindeutig definierten Variablen entstehen. Damit wird für die weitere Bearbeitung und Nutzung der Daten festgelegt, „welcher Eintrag in der Datenmatrix welche Information repräsentiert" (Lück/Landrock 2014: 399). Hierfür werden Regeln und Standards vereinbart, die die Daten nutzbar und ggf. vergleichbar machen und den Datenaufbereitungsprozess weniger fehleranfällig gestalten. Die Regeln können sich generell auf den Datenaufbereitungsworkflow beziehen bzw. unmittelbar auf die Struktur der Daten und Metadaten in einem Datensatz gerichtet sein.

Generelle Regeln definieren z.B. die Bearbeitungsstufen, die die Daten durchlaufen sollen. Diese können von der Überprüfung und Bearbeitung der Daten über ihre Anonymisierung bis hin zur Integration von Einzeldatensätzen reichen. Damit werden bereits die Grundlagen für den Datenaufbereitungsworkflow und die Schritte der Variablencodierung gelegt. Darüber hinaus kann vereinbart werden, dass die Analysefiles alle verfügbaren empirischen Informationen enthalten und der Grad der Informationstiefe der Daten möglichst dem der Rohdaten entspricht. Ist z.B. eine Kategorisierung von Informationen für Variablen wie Alter oder Einkommen erforderlich, werden neben den Variablen mit höherem Aggregationsniveau die detaillierten Quellvariablen in den Daten behalten und zugänglich gemacht. Enthalten die Daten potentiell datenschutzrelevante Informationen, wie z.B. regionale Kennziffern

oder detaillierte sozio-demographische Variablen, werden diese erst in den Analysefiles, die über die entsprechenden Publikationswege für Dritte zugänglich gemacht werden, gelöscht. Zusätzlich kann für die Datenfiles innerhalb eines Projekts oder zwischen Projekten vereinbart werden, die Codierung der Variablen sowie ihre Beschreibung mit Metadaten nach einheitlichen Regeln vorzunehmen. Die Anwendung standardisierter Schemata, wie z.B. eines Systems negativer Werte für die Definition der fehlenden Werte, vereinfacht erfahrungsgemäß den Aufbau und die Dokumentation der Daten, reduziert die Fehlerquellen während der Datenaufbereitung und erhöht die Transparenz für die Datennutzenden.

Konkrete Codierregeln werden vereinbart und angewendet, um die Struktur eines Datensatzes, seine Variablen und Werte sowie die beschreibenden Metadaten (Variablen- und Wertelabels) zu definieren. Hierfür werden bereits im Erhebungsinstrument den Fragen eindeutige Variablennamen und den Antwortkategorien der geschlossenen Fragen eindeutige Werte zugewiesen. Das Erhebungsinstrument bildet damit den Master für die Erstellung eines Setup File (oder Codeplans), auf dessen Grundlage die Struktur des Rohdatenfiles definiert werden kann. Bei den einzelnen Schritten der Variablencodierung kann dann von bestehenden und in der Projektpraxis erprobten Regeln ausgegangen werden. Sie helfen im unmittelbaren Projektkontext, geeignete Variablennamen abzuleiten, die Skalenniveaus und Formate zu definieren, aussagekräftige Variablen- und Wertelabels aufzubauen sowie ein passendes Codierschema für die fehlenden Werte zu bestimmen. Für detaillierte Zusammenstellungen entsprechender Regeln und Standards verweisen wir auf Netscher/Eder (2018); Ebel und Trixa (2015), Lück und Landrock (2014), Jensen (2012) sowie Survey Research Center (2016).

Abbildung 6.2: Schritte der Variablencodierung



Quelle: Eigene Darstellung

Ausgehend hiervon wird im Weiteren festgelegt, wann im Projektverlauf die Variablen zu generieren und welche Codierschritte hierfür erforderlich sind. Variablenübersichten in Excel-Format haben sich in der Projektpraxis hierfür bewährt. Sie unterstützen die Organisation dieses Schrittes und schaffen die erforderliche Transparenz im Projektverlauf. Abbildung 6.2 zeigt einen Ausschnitt aus einer Variablenübersicht, die für die European Values

Study (EVS) 2008[3] aufgebaut wurde. Das Beispiel geht von einem einfachen nationalen Datensatz aus und zeigt vier Bearbeitungsstufen eines Datensatzes sowie die jeweils enthaltenen bzw. aufzubauenden Variablen. Es beschreibt exemplarisch sechs Schritte der Variablencodierung (EVS 2016), die gleichermaßen auf andere Projekte anwendbar sind.

Das erste Beispiel in der Abbildung bezieht sich auf die inhaltlichen Variablen in einem Datensatz. Der überwiegende Teil dieser Variablen liegt bereits im Rohdatenfile in einer Form vor, die nachfolgende Datenanalysen unterstützt. Weitere Transformationsschritte können notwendig werden, wenn z.B. die fehlenden Werte erst nachträglich in ein einheitliches Schema codiert werden (Survey Research Center 2016: 637ff.).

Das zweite Beispiel beschreibt die Codierung soziodemographischer Variablen wie Beruf bzw. berufliche Tätigkeit oder Region. Werden die Antworten der Befragten als Textinformationen erhoben, müssen diese in nachfolgenden Schritten standardisiert werden. Für die Variable *Region* erfolgt dies bereits während der Datenerhebung, indem die Informationen in den *NUTS*-Standard der amtlichen Statistik für die Europäische Union (NUTS 2006) transformiert werden. Auf dieser Grundlage werden in späteren Codierschritten weitere Variablen auf höherem Aggregationsniveau generiert.

Die Beispiele drei und vier zeigen Schritte der Datenanonymisierung, die erforderlich werden, um eine Re-Identifizierung der Befragten zu verhindern. Wie in Kapitel 4.3 beschrieben, wird hierfür die Informationstiefe der Variablen *Beruf* und *Region* an die gewählten Publikationswege für die Daten angepasst. In Beispiel drei werden Variablen mit detaillierten Textinformationen aus den Analysefiles gelöscht. Ein geschützter Zugangsweg mit Datennutzungsvertrag kann dann gewählt werden, um die Daten mit den standardisierten, aber noch immer sehr detaillierten Angaben zu Beruf und Region bereitzustellen (*Scientific Use File*). Darauf aufbauend zeigt Beispiel vier, wie die Informationstiefe der Daten weiter verringert wird, um nur Informationen mit höherem Aggregationsniveau in dem Analysefile zu belassen. Das entstehende *Public Use File* kann dann über Retrieval-Systeme mit direkten Downloadmöglichkeiten für eine breite Nutzung zugänglich gemacht werden.

Beispiel fünf bezieht sich auf neu konstruierte und technische Variablen, die die Daten in eine einfacher nutzbare Form bringen sollen (Lück/Landrock 2014: 402f.). Hierzu gehören klassierte Variablen (z.B. für das *Alter*), dichotome Variablen, die für Fragen mit Mehrfachantworten generiert werden, sowie technische und administrative Variablen (wie Identifikations- oder Versionsnummern).

Beispiel sechs ist auf die Codierung länderspezifischer Fragen gerichtet, die erforderlich sind, um z.B. Bildungsniveaus in unterschiedlichen Schul- und Ausbildungssystemen oder Religionszugehörigkeit adäquat erfassen zu können. Für die Harmonisierung *der Bildungsvariable* wird hier die *International Standard Classification of Education (ISCED 2011)* angewendet und für die Religionsvariable ein kollektionsspezifischer Standard. Um die vorgenommenen Harmonisierungsschritte später replizieren zu können, werden ergänzend zu den harmonisierten Variablen standardisierte länderspezifische Quellvariablen zur Verfügung gestellt.

Eine Variablenübersicht, wie in Abbildung 6.2 dargestellt, kann zu Beginn eines Projekts aufgebaut und sollte im Projektverlauf als Arbeitsgrundlage kontinuierlich aktualisiert werden. Sie kann nach Projektabschluss zusammen mit dem Erhebungsinstrument sowie der Dokumentation der Codierschritte als eine zusätzliche Informationsquelle für Nachnutzende der Daten bereitgestellt werden.

3 Die European Values Study (EVS) ist eine transnationale empirische Langzeitstudie, die seit 1981 in einem Zyklus von neun Jahren durchgeführt wird. Der beschriebene Workflow wurde für die Daten der vierten Erhebungswelle, EVS 2008, aufgebaut und an die Erfordernisse der fünften Erhebungswelle, EVS 2017, angepasst.

6.4 Prüfung und Behandlung von Datenfehlern

Ausgehend von den definierten Datensätzen und den in ihnen enthaltenen Variablen können die erforderlichen Prozeduren für die Überprüfung der Daten und die Behandlung von Datenfehlern geplant werden. Typische Quellen für Datenfehler lassen sich in den verschiedenen Projektphasen, beginnend mit der Definition der zu befragenden Population und des Samples bis hin zur Analyse der Daten und der Interpretation der Ergebnisse, identifizieren (Schäfer et al. 2006: 1ff.). In den folgenden Ausführungen beschränken wir uns auf Datenfehler, die ihre Ursache im Erhebungsinstrument haben können, während der Datenerhebung oder -erfassung produziert wurden bzw. das Ergebnis des Datenaufbereitungsprozesses selbst sind (s. auch Braun 2014: 760; Pötschke 2010: 55). Der Prozess der Fehlersuche ist dabei vorrangig auf „unmögliche, unwahrscheinliche und widersprüchliche Werte" in den Daten (Lück 2014: 403) gerichtet, die die Analyseergebnisse verfälschen können.

Analog zur Variablencodierung werden auch hier Regeln vereinbart, die im Projektverlauf entscheiden lassen, welche Prüfprozeduren wann einzusetzen sind. Prüfprozeduren werden generell angewendet, um die Daten auf mögliche Fehlerarten hin zu überprüfen. Ihr wiederholter Einsatz wird erforderlich, um Fehler infolge vorgenommener Datenmodifikationen aufzudecken oder die Inhalte verschiedener Datensatzversionen zu vergleichen. Wurden potentielle Fehler in den Daten gefunden, müssen in weiteren Schritten die Quellen der Fehler analysiert und geeignete Maßnahmen der Fehlerbehandlung angewendet werden.

6.4.1 Schritt 1: Überprüfung der Daten und Metadaten

In diesem Schritt werden formale und logische Prüfprozeduren angewendet, die sich auf die einzelnen Variablen, auf Variablenkombinationen oder auf Datensätze beziehen können. In Schaukasten 6.1 sind häufig auftretende Fehler zusammengestellt, die erfahrungsgemäß Gegenstand der Datenüberprüfung sind (Survey Research Center 2016: 651ff.; Lück/Landrock 2014: 405ff.; Eurofound o.J.: 4f.). Die Variablen eines Datensatzes werden z.B. nach System Missing Values, Wild Codes oder Outliers untersucht. Verschiedene Kombinationen von Variablen werden gebildet, um Antwortmuster der Befragten auf Konsistenz und Plausibilität hin zu überprüfen. Auf der Ebene der Datensätze werden die erwarteten Fall- und Variablenzahlen bzw. die Struktur des Datensatzes kontrolliert.

Die Position der Prüfprozedur im Projektworkflow und die Art der Datenfehler entscheiden darüber, inwieweit visuelle Datenchecks mit einfachen oder komplexen Syntax-Checks verbunden werden können und welche Maßnahmen der Fehleranalyse erforderlich sind:

- *Visuelle Checks* werden vorgenommen, um die Übereinstimmung zwischen Erhebungsinstrument und Datensatz zu überprüfen, Variablen- und Wertelabels, Formate und Messniveaus der Variablen zu kontrollieren sowie Auffälligkeiten in den Daten zu identifizieren.
- *Einfache und komplexe Syntax-Checks* ermöglichen es, mit Hilfe von Häufigkeitsverteilungen, Kreuztabellen, deskriptiven Statistiken oder programmierten Prüfroutinen die Daten auf *benutzerdefinierte fehlende Werte*, *System Missing Values* und *Wild Codes* hin zu überprüfen. Ebenso können Filterinkonsistenzen, Inkonsistenzen innerhalb einer Item-Batterie bzw. logisch ausgeschlossene Beziehungen zwischen Variablen aufgedeckt sowie generell Datenmodifikationen überprüft werden.
- *Zusätzliche Kontextinformationen*, wie z.B. Bevölkerungsstatistiken, sind für Plausibilitätskontrollen soziodemographischer Verteilungen in den Daten sowie für die Überprüfung von Werten erforderlich, die außerhalb des realistischen Bereichs liegen. Sie unterstützen darüber hinaus die Kontrolle der Daten hinsichtlich sensibler Informationen über die befragte Person.

Schaukasten 6.1: Überblick über potentielle Datenfehler

Ebene	*Die Prüfung der Daten bezieht sich auf*
Variable	• nicht eindeutige IDs für Befragte, Interviewer, Datensätze • fehlerhafte Variablennamen und Werte sowie Variablenlabels/Werteetiketten • Fehler in Formaten, Messniveaus, Missing Value Definitionen • Wert außerhalb des realistischen Bereichs (Outliers) • Wert außerhalb des gültigen Bereichs (Wild codes) • ungültige fehlende Wert
befragte Person	• Inkonsistenzen im Antwortverhalten innerhalb des Fragebogens • Filterführungsfehler • duplizierte Fälle • unvollständige Interviews • Werte, die im Widerspruch zu Kontextinformationen oder Paradaten stehen • ungewöhnliche Antwortmuster
Datensatz	• fehlerhafte Anzahl der Variablen und/oder Fälle im Datensatz • Abweichungen in der Reihenfolge der Variablen im Datensatz und Fragebogen

Quelle: Eigene Darstellung

6.4.2 Schritt 2: Analyse der Quellen potentieller Datenfehler

Werden potentielle Fehler in den Daten identifiziert, müssen die möglichen Ursachen hierfür geklärt werden. Auf dieser Grundlage können geeignete Maßnahmen der Fehlerbehandlung gewählt werden. In Schaukasten 6.2 sind mögliche Quellen potentieller Datenfehler aufgeführt. Sie lassen sich auf die verschiedenen Phasen des Lebenszyklusmodells zurückführen und können das Erhebungsinstrument und den Übersetzungsprozess, die Befragungssituation und Datenerfassung sowie die Datenaufbereitung betreffen.

Schaukasten 6.2: Mögliche Quellen potentieller Datenfehler in den Projektphasen

Projektphase	*Mögliche Quellen für Datenfehler*
Planungsphase	Erhebungsinstrument: fehlende/sich überschneidende Antwortkategorien, komplexe oder fehlerhafte Filterführung
	Übersetzungsprozess: fehlerhafte Übersetzung, nicht vergleichbare Fragestellung, gedrehte Skalen, fehlende/zusätzliche Antwortkategorien
Datenerhebung/ Datenerfassung	Befragte oder Interviewer: unzutreffende bzw. bewusst falsche Angaben
	Interviewer: bewusste Fälschung bzw. Duplikation von Fällen
	Codierer: fehlerhafte Codierung offener Angaben
	Datenerfassung: technische und manuelle Fehler, mehrfache Erfassung von Fällen
Datenaufbereitung	• Fehlerhafte Codierung während der Fehlersuche und Fehlerbehandlung • der Standardisierung und Harmonisierung • der Generierung neuer Variablen • der Integration/Kumulation der Daten • der Anonymisierung der Daten

Quelle: Eigene Darstellung

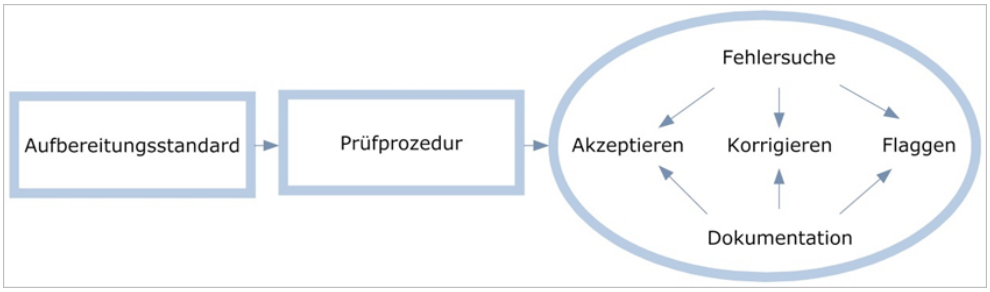
Der Aufwand für den Schritt der Fehlersuche und Fehlerbehandlung lässt sich verringern, wenn z.B. das Erhebungsinstrument getestet wurde. Ist es zudem möglich, die Daten zeitnah zu ihrer Entstehung zu überprüfen, kann häufig auf die Quelle der Informationen zurückgegriffen werden (Lück/Landrock 2014: 406f.). Das ermöglicht es erfahrungsgemäß, Fehler mit geringerem Aufwand aufzuklären und Informationsverluste in den Daten zu vermeiden. Dementsprechend werden in *computer-assisted*-Interviews Prüfprozeduren direkt in den digital programmierten Fragebogen implementiert. Bei unzulässigen bzw. inkonsistenten Antworten werden den interviewenden bzw. befragten Personen Fehlermeldungen angezeigt, die sie bereits während der Befragungssituation überprüfen und ggf. korrigieren können (Survey Research Center 2016; Eurofound and GfK EU3C o.J.: 50ff.). Knüpft bei *paper-and-pencil*-Interviews oder schriftlichen Befragungen die Datenüberprüfung zeitlich an die Datenerhebung oder Datenerfassung an, kann ggf. auf die Erhebungsinstrumente zurückgegriffen werden, um z.B. fehlende Werte sowie Werte, die außerhalb des gültigen bzw. realistischen Wertebereichs liegen, zu überprüfen. Werden dagegen potentielle Datenfehler erst in einer zeitlich nachgelagerten Phase der Datenaufbereitung entdeckt, lassen sie sich z.T. nur noch schwer aufklären.

Prüfprozeduren sollten demzufolge, und sofern die Projektressourcen es erlauben, möglichst dann implementiert werden, wenn die originalen Erhebungsinstrumente bzw. die befragten, interviewenden, codierenden bzw. aufbereitenden Personen noch verfügbar sind (Survey Research Center 2016: 636).

6.4.3 Schritt 3: Behandlung und Dokumentation der identifizierten Datenfehler

Wurden potentielle Fehler in den Daten identifiziert, werden die vereinbarten Regeln der Fehlerbehandlung angewendet. Wie in Abbildung 6.3 dargestellt, können potentielle Datenfehler nach ihrer Überprüfung als korrekt akzeptiert, in einen anderen Wert recodiert oder *geflaggt* werden. *Flag*-Variablen zeigen das Vorliegen (Wert = 1) bzw. das Nicht-Vorliegen (Wert = 0) eines potentiellen Fehlers an (Kveder/Galico 2008: 6f.; Kampmann et al. 2014). Dadurch wird die Entscheidung, wie konkret mit dem jeweiligen Fehler in der Datenanalyse umgegangen werden soll, den Datennutzenden überlassen.

Abbildung 6.3: Möglichkeiten der Behandlung potentieller Datenfehler



Quelle: Eigene Darstellung (siehe auch Kampmann et al. 2014)

Die Regeln für die Behandlung potentieller Fehler werden projektspezifisch definiert. Nach Lück und Landrock (2014: 406) sollte hierbei zwischen *zweifelsfrei* und *nicht zweifelsfrei* nachweisbar fehlerhaften Werten unterschieden werden:

- Zweifelsfrei fehlerhafte Werte sollten korrigiert werden, wenn der wahre Wert mit hoher Wahrscheinlichkeit nachgewiesen werden kann.

- Zweifelsfrei fehlerhafte Werte sollten gelöscht bzw. durch einen *Missing Value* ersetzt werden, wenn der wahre Wert nicht mit hoher Wahrscheinlichkeit nachgewiesen werden kann.
- Kann für einen Wert nicht zweifelsfrei nachgewiesen werden, dass er fehlerhaft ist, sollte er in dem Datensatz mit Hilfe von *Flag*-Variablen markiert werden.

Ob Fehler in den Daten jedoch nur dokumentiert oder auch behandelt werden, kann z.B. von der Zahl der betroffenen Fälle abhängig gemacht werden. Sind mehr als nur einzelne Fälle betroffen, wird das Datenproblem dokumentiert und ggf. *geflaggt*, um damit die ursprüngliche Qualität der erhobenen Daten zu erhalten und sichtbar zu machen (Kolsrud et al. 2010: 64). So wird bei der European Value Study beispielsweise *geflaggt*, sobald (mehr als) 1 % der Stichprobe betroffen sind. Schaukasten 6.3 enthält einzelne Beispiele für Codierregeln, die für diese Datenkollektion aufgestellt wurden (EVS 2017). Zusammenstellungen über verschiedene Prüfverfahren und Entscheidungsregeln finden sich auch bei Lück und Landrock (2014: 407) und Jensen (2012).

Schaukasten 6.3: Beispiele für Codierregeln in einer multinationalen Umfrage

Univariate Codier-Regeln	Recodieren in ‚keine Angabe' und dokumentieren • Wert außerhalb des gültigen Bereichs (Wild codes) • Wert außerhalb des realistischen Bereichs (Outliers)
Multivariate Codier-Regeln	Dokumentieren • widersprüchliche Antworten auf verschiedene Fragen • nicht-plausible, aber logisch mögliche Werte des Befragten
Regeln für Filterfragen	Rekonstruieren/recodieren und dokumentieren • Missing Values nach Filterfragen werden recodiert: in trifft nicht zu, wenn der Befragte nicht gefragt werden sollte; • in keine Angabe, wenn der Befragte gefragt werden sollte

Quelle: Eigene Darstellung

Die Dokumentation des Prozesses der Fehlersuche und Fehlerbehandlung hat Bedeutung sowohl für Forschende, die die Daten erheben, als auch für diejenigen, die sie nachnutzen wollen. Primärforschenden hilft sie, ähnliche Fehlerquellen in nachfolgenden Erhebungen zu vermeiden (Survey Research Center 2016: 651f.). Für Sekundärforschende schafft sie Transparenz über die Qualität der vorliegenden Daten und ermöglicht, Datenmodifikationen nachzuvollziehen bzw. die Daten gezielt nach weiteren potentiellen Fehlern zu untersuchen (vgl. dazu Kapitel 8.1). Die Sicherung der Rohdaten bzw. die Bereitstellung eines Datensatzes, in dem die identifizierten Datenfehler markiert sind, erlaubt darüber hinaus, den potentiellen Wert des Datenaufbereitungsprozesses zu überprüfen (Kveder/Galico 2008: 3ff.).

6.5 Entwicklung eines Workflows der Datenaufbereitung

In den bisherigen Ausführungen wurden die im Projektverlauf zu generierenden Datensätze geplant, die Variablen definiert und hieraus Regeln und Prozeduren für die Variablencodierung sowie die Datenüberprüfung abgeleitet. Im Folgenden wird ein Workflow vorgestellt, der Forschenden helfen soll, die einzelnen Schritte in die Projektpraxis umzusetzen. Der Workflow in Abbildung 6.4 wurde für die Aufbereitung eines einfachen Datenfiles entwickelt. Er basiert auf einem System von Datenfiles, Syntax-Files und Ordnern und lässt sich an unterschiedliche Projekterfordernisse anpassen. So können kleinere Projekte ggf. weniger

Prüfschritte benötigen und müssen komparative Projekte weitere Schritte der Harmonisierung und Integration der Daten einbauen. Beispiele hierzu finden sich bei Harzenetter und Wronski (2015: 10ff.), Kampmann et al. (2014) sowie bei Brislinger et al. (2011). Der entstehende Datenaufbereitungsworkflow setzt sich aus drei Hauptbausteinen zusammen:

1. eine Ordnerstruktur, die eine strukturierte Ablage der Daten und Dokumentationen ermöglicht,
2. Datenfiles, die die einzelnen Schritte der Erfassung, Bearbeitung und Publikation der Daten sichtbar machen, sowie
3. Syntax-Files, die die geplanten Schritte der Variablencodierung, Fehlersuche und Fehlerbehandlung ausführen.

Ausgehend von dem in Abbildung 6.1 beschriebenen Lebenszyklusmodell beginnt der Workflow mit den Projektempfehlungen aus der Phase der Projektplanung, beinhaltet die Rohdaten und Informationen, die aus der Datenerhebung übergeben werden, und endet mit den Analysefiles, die für die sekundäranalytische Nutzung über ein Repositorium oder Datenarchiv bereitgestellt werden (s. Kapitel 7.4). Hierbei werden, wie oben ausgeführt, die Dateien in einer Ordnerstruktur abgelegt, definieren die Datenfiles die Struktur des Datenaufbereitungsworkflows und schaffen die Syntax-Files Transparenz über die einzelnen Bearbeitungsschritte.

6.5.1 Aufbau der Ordnerstruktur

Abbildung 6.4: Aufbau eines Datenaufbereitungsworkflows

Ordner für Dokumente aus der Planungsphase

- Projektempfehlungen
- Erhebungsinstrument

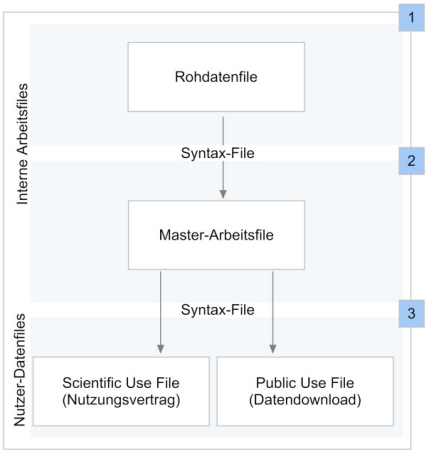
Ordner für Lieferung aus der Datenerhebung

- eingesetztes Erhebungsinstrument
- Rohdaten, Paradaten
- Feld- oder Methodenbericht

Kommentierte Syntax

- Erstellung des Datenfiles: FILE-NAMEN
- Rohdatenfile: FILE-NAMEN
- Quelldokumente: FILE-NAMEN
- Setup Datum: YYMMTT
- Bearbeiter: NAME
- Speicherort: PFAD-NAME
- Version History: Veränderungen zwischen Versionen

Ordner für den Aufbereitungsworkflow



Internes Rohdatenfile (Schritt 1)

- Sicherung des Rohdatenfiles vor Überschreiben
- Überprüfung: Anzahl der Fälle und Variablen

Erstellung des internen Master-Arbeitsfiles (Schritt 2)

- visuelle Checks: Übereinstimmung Erhebungsinstrument und Daten, Formate, Labels, fehlende Variablen/Werte
- Konstruktion administrativer/technischer Variablen
- Definieren der Variablenlisten für Prüfprozeduren
- Konsistenzprüfungen
  - doppelte Befragten-Nummern
  - Wild Codes, Outliers
  - Filterführung, Item-Batterien und Mehrfachnennungen
  - Gewichtungsvariablen
- Überprüfung der fehlenden Werte
  - identifizieren/recodieren der System Missings
  - standardisieren der Werte und Wertelabels
- Datensatzstruktur: Variablen Reihenfolge, Formate, Messniveaus

Erstellung der Analysefiles für die Forschung (Schritt 3)

- Umsetzung des Anonymisierungskonzepts für Publikationsweg: Nutzungsvertrag oder Datendownload
- abschließende Überprüfung der Analysefiles:
  - Anzahl der Fälle und Variablen, fehlende Werte, Labels

Quelle: Eigene Darstellung in Anlehnung an Brislinger et al. (2011)

Datei-Systeme mit einer bereits zu Beginn des Projekts definierten Ordnerstruktur vereinfachen erfahrungsgemäß die Ablage und das Auffinden der Daten und Dokumentationen, schaffen Transparenz und helfen Informationsverluste zu vermeiden. Wie in Kapitel 5 beschrieben kann eine Ordnerstruktur, die entlang des Lebenszyklus eines Projekts aufgebaut wird, am ehesten den Erfordernissen des Daten- und Informationsmanagements gerecht werden. Dem folgend sind in Abbildung 6.4 exemplarisch Ordner für die Projektempfehlungen aus der *Planungsphase* sowie die Daten und Dokumentationen aus der *Datenerhebungsphase* angelegt. Der Ordner *Aufbereitungsworkflow* enthält die Daten und Informationen, die während der Datenaufbereitung selbst bearbeitet oder neu erstellt werden. Seine weitere Untergliederung in Unterordner ermöglicht es, die generierten Datenfiles strukturiert abzulegen. Das Dateisystem, das hierbei schrittweise entsteht, macht die produzierten Daten, Dokumentationen und Prozessinformationen für alle Projektbeteiligten sichtbar und zugänglich. Am Ende eines Projekts bildet es eine gute Grundlage für die Auswahl der Dateien, die im Projektkontext aufzubewahren sind bzw. an ein Datenarchiv oder Repository übergeben werden müssen.

6.5.2 Aufbau der Datenfiles

Wie bereits ausgeführt, können die im Datenaufbereitungsprozess zu generierenden Datenfiles nach ihren Bearbeitungsstufen und Aufgaben bei der Realisierung der Projektziele unterschieden werden. Die *Rohdatenfiles* aus der Datenerhebung werden gesichert und schaffen Transparenz über den Datenaufbereitungsprozess, indem vorgenommene Datenmodifikationen jederzeit zurückverfolgt bzw. verloren gegangene Daten rekonstruiert werden können. Sie bilden die Grundlage für die internen Master-Arbeitsfiles, die alle erhobenen und neu generierten Variablen enthalten und den Ausgangspunkt für spätere Daten-Updates sowie die Dokumentation der Daten bilden. Aus diesen Datenfiles wiederum werden am Ende des Workflows die Analysefiles generiert, die für die Forschung im Projekt sowie darüber hinausgehend für eine breitere Nachnutzung der Daten über verschiedene Publikationswege bereitgestellt werden können.

6.5.3 Aufbau der Syntax-Files

Mit Hilfe kommentierter Syntax-Files werden die beschriebenen Schritte der Variablencodierung sowie der Fehlersuche und Fehlerbehandlung durchgeführt. Sie sollten möglichst strukturiert aufgebaut sein und, wie in Abbildung 6.4 dargestellt, z.B. Informationen über den Bearbeitenden, den Zeitpunkt der Bearbeitung und den zu bearbeitenden Datensatz enthalten. Ein Verzeichnis der Bearbeitungsschritte sowie ihre Kommentierung ermöglichen es auch Dritten, die Programmlogik nachzuvollziehen. Mit Hilfe der Syntax-Files werden die Rohdaten in Master-Arbeitsfiles und diese wiederum in Analysefiles transformiert. Hierbei werden, wie in Abbildung 6.4 gezeigt, die Rohdaten aufgerufen (Schritt 1), schrittweise überprüft, korrigiert und standardisiert und als internes Master-Arbeitsfile gesichert (Schritt 2). In einem weiteren Schritt werden die Anonymisierungsmaßnahmen umgesetzt, um die entstehenden Analysefiles über die gewünschten Publikationswege zur Verfügung stellen zu können (Schritt 3).

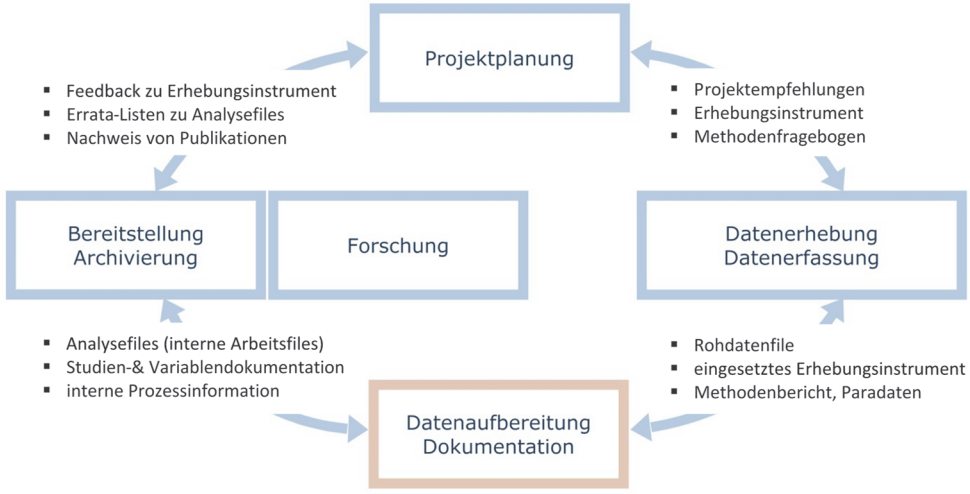
Wird ein solches Modell bereits zu Beginn des Projekts geplant, macht es den Datenaufbereitungsworkflow für alle Beteiligten transparent und verständlich. Es hilft darüber hinaus, die eingangs formulierte Frage zu beantworten, wann im Datenaufbereitungsworkflow die

erforderlichen Bearbeitungsschritte implementiert werden sollten und wo die hierfür notwendigen und die produzierten Dateien abgelegt bzw. abzulegen sind.

6.6 Informationstransfer und Datendokumentation

Die Planungsdokumente, Erhebungsinstrumente bzw. internen Prozessinformationen, die im Projektverlauf aufgebaut werden, bilden die Grundlage für die Datendokumentation. Diese umfasst im idealen Falle den gesamten Prozess der Planung, Erhebung und Aufbereitung der Daten (Häder 2006). Der Transfer dieser Informationen zwischen den einzelnen Phasen schafft die erforderliche Transparenz und bildet die Grundlage für das Zusammenwirken der Projektmitglieder sowie die Entscheidungen, die sie im Umgang mit den Daten treffen müssen. Abbildung 6.5 zeigt das eingangs verwendete Lebenszyklusmodell, diesmal erweitert um Informationen, die in den einzelnen Projektphasen generiert und an die nachfolgenden Phasen übergeben werden.

Abbildung 6.5: Informationstransfer zwischen den Projektphasen



Quelle: Eigene Darstellung in Anlehnung an Survey Research Center (2016)

Für eine übersichtliche Darstellung wird hier nur eine Auswahl an möglichen Files aufgeführt und ist der Transfer zwischen den Phasen auf die jeweils neu entstehenden Daten und Informationen begrenzt. In der Projektpraxis müssen jedoch Dokumentationen, wie z.B. die Projektempfehlungen, nicht nur an die unmittelbar nachfolgende Phase weitergegeben werden, sondern allen Projektmitgliedern in der aktuellen Version jederzeit zur Verfügung stehen.

Schaut man aus Sicht der *Datenaufbereitung und Dokumentation* auf das Lebenszyklusmodell, dann werden die Daten und Informationen der vorausgehenden Projektphasen an diese Phase transferiert und bilden die Grundlage für die weiteren Aufbereitungsschritte. Hierbei werden neue Informationen generiert, die die vorgenommenen Datenmodifikationen und die Besonderheiten in den Daten beschreiben. Im Ergebnis entsteht eine Datendokumentation, die im idealen Fall die wichtigen im Projektverlauf produzierten Informationen umfasst.

Sollen die Daten nach Abschluss des Projektes für eine nachhaltige Sicherung und weitere Nutzung an ein Datenarchiv oder Repositorium übergeben werden, muss eine Auswahl aus dem umfangreichen Informationsbestand getroffen werden. Die entstehende Datendokumentation sollte Informationen auf Projekt-, Studien- und Variablenebene enthalten und damit sowohl den Kontext des Projekts als auch den unmittelbaren Entstehungs- und Bearbeitungsprozess der Daten für Dritte verständlich machen. Hauptbestandteile sind: (1) die Projektempfehlungen, das Erhebungsinstrument sowie Interviewer-Informationen aus der Planungsphase, die helfen sollen, die Projektziele zu erreichen und Daten mit der angestrebten Qualität zu produzieren, (2) Informationen, die während der Erhebung der Daten generiert wurden und Auskunft über z.B. die Antwortbereitschaft und das Antwortverhalten der befragten Personen geben, und (3) interne Prozessinformationen, wie die Arbeitsfiles und Syntaxfiles, die es darüber hinaus ermöglichen, Datenprobleme, die während späterer Analysen gefunden werden, aufzuklären. Sie bilden zusammen mit der Beschreibung des Datenaufbereitungsworkflows einen guten Ausgangspunkt, falls nachfolgende Erhebungswellen geplant sind (Survey Research Center 2016: 667).

Ein Teil dieser Informationen kann in einem weiteren Schritt in einem Methodenbericht sowie Variablenreport aggregiert werden, die Forschenden den Zugang zu und den Umgang mit den Daten erleichtern. Sie geben am Ende eines Projekts detailliert Auskunft über die Entstehung der Daten, ihren Inhalt und die Qualität und ermöglichen damit eine adäquate Nutzung (Lück/Landrock 2014).

6.6.1 Methodenbericht

Der Methodenbericht beinhaltet neben allgemeinen projektbeschreibenden Informationen – wie Titel und Zitation der Daten, Primärforscher/innen sowie Quellen der Projektfinanzierung – methodische Informationen aus den einzelnen Projektphasen. Er beschreibt den Prozess der Fragebogenentwicklung, die Grundgesamtheit, das Auswahlverfahren und enthält Informationen über den Befragungszeitraum und die Befragungsmethode sowie die Ausschöpfung der Stichprobe. Projekte, die die Daten selbst erheben, müssen die erforderlichen Informationen möglichst zeitnah zu ihrer Entstehung und umfassend protokollieren. Wird ein Erhebungsinstitut mit der Datenerhebung beauftragt, ist der Methodenbericht oder Feldbericht zumeist Bestandteil der Datenlieferung an das Projekt. Bei multinationalen Projekten basiert er auf einem Methoden- oder Feldfragebogen, der Teil der Planungsdokumente ist und es ermöglicht, auch bei mehreren involvierten Erhebungsinstituten die Entstehung der Daten strukturiert zu beschreiben.

6.6.2 Variablenreport

Der Variablenreport beschreibt die in dem Datensatz enthaltenen Variablen und gibt einen Überblick über die eingesetzten Prüfprozeduren und vorgenommenen Datenmodifikationen. Die Dokumentation der einzelnen Bearbeitungsschritte macht es auch Forschenden außerhalb des Projekts möglich, die Qualität der Daten zu bewerten und zu entscheiden, ob im eigenen Forschungskontext weitere Prüfprozeduren bzw. Datentransformationen erforderlich sind. Hierfür werden Informationen aus den verschiedenen Phasen des Forschungsprojekts und aus unterschiedlichen Quellen zusammengeführt. Im idealen Falle werden die einzelnen Variablen durch die Metadaten des Datensatzes (Variablenname, Werte sowie Variablen- und Wertelabels) beschrieben und mit weiteren Informationen verknüpft. Das sind häufig der exakte Wortlaut der Frage aus dem Erhebungsinstrument sowie die absoluten und

relativen Häufigkeitsverteilungen des Analysefiles, die einen ersten Blick auf die Verteilungen in den Daten geben. Darüber hinaus können Kommentare zu den Variablen ergänzt werden, die z.B. über die Quelle der eingesetzten Frage, die verwendeten Standards sowie Besonderheiten in den Daten informieren. Für den Aufbau solcher Variablendokumentationen stehen Tools zur Verfügung, die internationale Metadatenstandards anwenden und damit einen plattformübergreifenden Transfer der Dokumentationen erlauben, wie im zweiten Teil von Kapitel 9 näher beschrieben wird. Nach Abschluss des Projekts kann die Variablendokumentation über Retrieval-Systeme zugänglich gemacht werden und Datennutzende bei Recherchen in den Fragetexten und damit bei der direkten Erschließung der Dateninhalte unterstützen.

Die Dokumentationen, die zusammen mit den Daten weitergegeben werden, bilden im Kontext eines Datenarchivs oder Repositoriums eine wichtige Grundlage für die Erstellung strukturierter Metadaten. Die Aufgabe der Metadaten ist es dann, die Daten inhaltlich und methodisch möglichst exakt und umfassend zu beschreiben (vgl. Kapitel 9.2) und in Retrieval-Systemen, wie z.B. Datenkataloge, für Sekundärforschende einfach auffindbar und zugänglich zu machen (vgl. Kapitel 8.1).

6.7 Fazit

Eine systematische Dokumentation der Entstehung und Bearbeitung der Daten gibt im unmittelbaren Projektkontext einen guten Einblick in die Datenqualität und schafft einen adäquaten Datenzugang. Vermittelt über ein Repositorium oder Datenarchiv ermöglicht sie es auch Dritten, die Daten im Rahmen eigener Forschungsprojekte zu verwenden sowie die originären Projektergebnisse zu replizieren. Eine nachhaltige Sicherung der Ergebnisse des Projekts erlaubt es darüber hinaus, diese als einen Ausgangspunkt für die zukünftige Projektplanung zu nutzen.

Die Grundlage hierfür ist die Planung und Organisation der Datenaufbereitung und Dokumentation im Projektverlauf, wofür sich Lebenszyklusmodelle als sehr hilfreich erweisen. Sie schaffen die erforderliche Transparenz sowohl für die einzelnen Arbeitsschritte als auch für komplexe Aufbereitungsworkflows und ermöglichen es den Forschenden, adäquate Wege für die Realisierung der Projektziele zu finden und dabei die Projektressourcen zu berücksichtigen.

Literaturverzeichnis

Ball, Alex (2012): Review of Data Management Lifecycle Models. Version 1.0. REDm-MED Project Document. University of Bath. http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf [Zugriff: 02.05.2018].

Braun, Michael (2014): Interkulturell vergleichende Umfragen. In: Nina Baur, Jörg Blasius (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer, S. 757-766.

Brislinger, Evelyn/Nijs Bik, Emile de/Harzenetter, Karoline/Hauser, Kristina/Kampmann, Jara/Kurti, Dafina/Luijkx, Ruud/Ortmanns, Verena/Rokven, Josja/Sieben, Inge/Solanes Ros, Ivet/Stam, Kirsten/Weijer, Steve van de/Vlimmeren Eva van/Zenk-Möltgen, Wolfgang (2011): European Values Study 2008: Project and Data Management. GESIS Technical Reports 2011/14. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2011/TechnicalReport_2011-14.pdf [Zugriff: 23.06.2018].

DFG – Deutsche Forschungsgemeinschaft (2015): Umgang mit Forschungsdaten, Leitlinien zum Umgang mit Forschungsdaten.

http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf [Zugriff: 23.06.2018].

Ebel, Thomas/Trixa, Jessica (2015): Hinweise zur Aufbereitung quantitativer Daten. GESIS Papers 2015/09. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-432235 [Zugriff: 25.06.2018].

Eurofound (o.J.): 3rd European Quality of Life Survey. Data Editing & Cleaning Report. EU27 and non-EU Countries (Internal Report). UK Data Archive Study Number 7316 – European Quality of Life Survey, 2011–2012. http://doc.ukdataservice.ac.uk/doc/7316/mrdoc/pdf/7316_3rd_eqls_data_editing_and_cleaning_report.pdf [Zugriff: 10.07.2018].

Eurofound and GfK EU3C (o.J.): 3rd European Quality of Life Survey. Technical Report. Working Document for The European Foundation for the Improvement of Living and Working Conditions. https://www.eurofound.europa.eu/sites/default/files/ef_files/surveys/eqls/2011/documents/technicalreport.pdf [Zugriff: 10.07.2018].

EVS – European Values Study (2016): EVS 2008 - Variable Report Integrated Dataset. GESIS-Variable Report 2016/2. https://dbk.gesis.org/dbksearch/file.asp?file=ZA4800_cdb.pdf [Zugriff: 25.06.2018].

EVS – European Values Study (2017): Data Processing Guidelines for the European Values Study 2017. GESIS-DAS & EVS Tilburg University. Internes Arbeitspapier. Mai 2018.

Felderer, Barbara/Birg, Alexandra/Kreuter, Frauke (2014): Paradaten. In: Baur, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer, S. 357-365.

GESIS – Leibniz-Institut für Sozialwissenschaften (2016): GESIS Survey Guidelines. http://www.gesis.org/gesis-survey-guidelines/home/ [Zugriff: 01.06.2018].

Harzenetter, Karoline/Wronski, Pamela (2015): Aufbereitung und Dokumentation von Studien der Bundeszentrale für gesundheitliche Aufklärung (BZgA). GESIS-DAS. Interner Abschlussbericht. 29.09.2015.

Häder, Michael (2006). Empirische Sozialforschung. Eine Einführung. Wiesbaden: VS.

Jensen, Uwe (2011): Datenmanagementpläne. In: Büttner, Stephan/Hobohm, Hans-Christoph/Müller, Lars (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock + Herchen, S. 71-82.

Jensen, Uwe (2012): Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten. GESIS Technical Reports 2012/07. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-320650 [Zugriff: 02. 05. 2018].

Kampmann, Jara/Harzenetter, Karoline/Wronski, Pamela/Brislinger, Evelyn/Solanes Ros, Ivet (2014): Data Processing Toolkit: Help, Tools, and Features. GESIS-DAS. Internes Arbeitspapier.

Kolsrud, Kirstine/Midtsæter, Hege/Orten, Hilde/ Skjåk, Knut Kalgraff/Øvrebø, Ole-Petter (2010): Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services. In: Ask: Research and Methods, Vol. 19 (1, 2010), S. 51-92. http://hdl.handle.net/1811/69570 [Zugriff: 10.07.2018]

Kveder, Andrej/Galico, Alexandra (2008): Guidelines for Cleaning and Harmonization of Generations and Gender Survey Data. https://www.ggp-i.org/sites/default/files/questionnaires/GGP_2008_DCHGuide_1.pdf [Zugriff: 23.06.2018].

Netscher, Sebastian/Eder, Christina (Hrsg.) (2018): Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research. GESIS Papers, 2018/22. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59492-3 [Zugriff 26.10.2018].

NUTS – Nomenclature of Units for Territorial Statistics (2006): Eurostats. Regions in the European Union. Nomenclature of Territorial Units for Statistics. NUTS 2006/EU-27. 17.01.2008. http://ec.europa.eu/eurostat/web/nuts/publications [Zugriff: 23.06.2018].

Pötschke, Manuela (2010): Datengewinnung und Datenaufbereitung. In: Wolf, Christof/Best, Henning (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS, S. 41-64.

Schäfer, Christin/Bömermann, Hartmut/Nauenburg, Ricarda/Wenzel, Karsten (2006): Data Driven Identification of Sources of Errors for Improving Survey Quality. Proceedings of Q2006. European Conference on Quality in Survey Statistics. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.610.3024&rep=rep1&type=pdf [Zugriff: 23.06.2018].

Survey Research Center (2016): Guidelines for Best Practice in Cross-cultural Surveys. Full Guidelines. Ann Arbor, MI: Institute for Social Research, University of Michigan. http://ccsg.isr.umich.edu/ [Zugriff: 02.05.2018].

Lück, Detlev/Landrock, Uta (2014): Datenaufbereitung und Datenbereinigung in der quantitativen Sozialforschung. In: Baur, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer, S. 397-409.

Linkverzeichnis

ESS – European Social Survey: http://www.europeansocialsurvey.org/data/round-index.html [Zugriff: 20.05.2018].
EVS – European Values Study (2017): https://europeanvaluesstudy.eu/methodology-data-documentation/survey-2017/ [Zugriff: 12.11.2018]
EVS - European Values Study (2008): https://europeanvaluesstudy.eu/methodology-data-documentation/previous-surveys-1981-2008/survey-2008/ [Zugriff: 12.11.2018]
ISCED (2011): International Standard Classification of Education. UNESCO Institute for Statistics: http://uis.unesco.org/en/topic/international-standard-classification-education-isced [Zugriff: 12.11.2018]